Benford's Law and why the integers are not what we think they are: A critical numeracy of Benford's Law

Rex Stoessiger
Quality Educational Services, NSW
<rex_stoessiger@southcom.com.au>

Introduction

When we examine numbers in the newspaper or magazines we might expect that their first digits are just as likely to be 8 or 9 as a 1 or a 2 and we might assume that each of the nine digits (zero is not used as a first digit of course) will occur $\frac{1}{9}$ of the time. Unexpectedly this turns out to be untrue in many situations.

It was the astronomer Newcomb (1881) who first noticed that the leading digits of numbers are not randomly distributed but that the smaller digits, particularly 1, were much more common. Newcomb did not develop the law from any insight into the number system; it came to him from observing how numbers seem to be used in life. He simply noticed that the first pages of log tables, which show numbers starting with 1, were much grubbier than pages starting with other digits. He realised that the log table users, he and his fellow astronomers, must be looking up the logs of numbers that started with the digit 1 much more commonly than other digits.

Table 1. Predicted probabilities of leading digits from Benford's Law.

Leading digit	Probability p (Benford's Law)
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046
Total	1.000

Why were Newcomb's colleagues so often looking up numbers which begin with digit 1 and less for the other digits? He formulated the expression given below (stating it was self evident) to show the frequency of first digits. The result was surprising both to Newcomb and his colleagues. So surprising that the finding was promptly forgotten! That the digit 1 occurs most often as the first digit is still surprising to most people today.

The law was rediscovered by Benford (1938) from the same observation of dirty log table pages in 1938 and states that, in certain circumstances, the distribution of the leading digit, D, in a collection of numerical data is given by the logarithmic formula,

$p = \log_{10}\left(1+\frac{1}{D}\right)$ for D = {1,2, … ,9} where p is the probability of digit D.

This is now usually called Benford's Law. The frequencies of the digits given by the Law are shown in Table 1. It shows that digit 1 occurs first 30% of the time while digit 9 occurs as first digit only 4.6% of the time. This is a six-fold difference and it would be expected that we would have noticed this about our world but it seems that most people are completely unaware of the disparity.

Benford made a wide examination of different situations where the Law might be valid and was able to show that it seemed to apply not just to scientific data but to tables of populations, river areas, numbers in newspapers and the Readers Digest, tables of physical constants, addresses, and cost data (see Weisstein, 2010, for Benford's tables).

Benford's Law requires that either 1 or 2 should be the leading digit about 47.7% of the time while 8 and 9 should be the initial digits on less than 10% of occasions. For most of us this is not an intuitive result. Without more information most people assume that the nine leading digits should be randomly distributed and the smaller digits should not be privileged in any way. What is special about 1 and 2 that one or the other should occur as the first digit nearly half the time?

Benford's tables and more recent research shows that Benford's Law applies to a surprisingly large variety of data although it does not apply to normally distributed data (where the numbers cluster about a mean and hence many have the same first digit), random numbers such as lottery results where the first digits are truly random and telephone numbers.

The more general form of the Law predicts that the frequencies of multiple digits, *dd*, are of the same form,

$$p = \log_{10}\ 1+\frac{1}{dd}$$

where *dd* is any string of digits in {1, 2, … 9}. For example, the probability that 2 is first followed next by 5 then by 3 is

$$p(253) = \log_{10}\ 1+\frac{1}{253}\ = 0.00171$$

Benford's Law is scale (and hence units) independent and the probabilities for the nine digits always sum to 1.0 as required by probability theory (zero

is excluded from use as a first digit but can occur as a subsequent digit). The Law is not peculiar to base 10 numbers but applies to numbers in any base, b, and can be expressed as

$$p = \log_b \left(1 + \frac{1}{dd}\right)$$

Anomalous but ubiquitous

Benford was so surprised with the distribution that he called his discovery the 'law of anomalous numbers'. In other words 'proper' numbers should have the first digits randomly distributed with a 1 just as likely as an 8 or a 9—each occurring as the first digit $\frac{1}{9}$ of the time. Numbers which did not behave like this he regarded as anomalous. This was despite his investigations that showed a large amount and a surprising variety of data behaved in this way.

That the numbers which obey Benford's Law are in some way strange has been a theme of papers published about the Law. Ley (1996) refers to the "peculiar distribution of the US stock indices' digits" and refers to numbers obeying Benford's Law as "the family of *anomalous* or *outlaw* numbers". Kafri (2009) states, "Benford's distribution of digits is counterintuitive". Fewster (2009) regards the findings as "counter-intuitive". Raimi (1976) has described it as "so common yet so surprising at first glance". Mathews (1999) describes it as: "It is a law so unexpected that at first many people simply refuse to believe it is true." Even Ted Hill, who has probably contributed as much as anyone to our understanding of the law, is quoted as saying, "For me the law is a prime example of a mathematical idea which is a surprise to everyone—even the experts" (Mathews, 1999, p. 28).

Yet an enormous amount of the data in our world conforms to Benford's Law. How is it possible that something so ubiquitous in everyday life could be seen as anomalous?

Critical numeracy

We can use the idea of critical numeracy to further examine Benford's Law. The next section will outline an understanding of critical numeracy as the first step.

Numeracy

While some people may have wished to define numeracy simply as arithmetic this has not happened. The role of wide areas of mathematics in everyday life has been well appreciated and the ability, for example, to use and understand graphs is seen by some as just as fundamental to numeracy as is arithmetic. The result has been definitions of numeracy that emphasise the practical or everyday uses of mathematics such as: "To be numerate is to have and be able

to use appropriate mathematical knowledge, understanding, skills, intuition and experience whenever they are needed in everyday life" (Education Department of Tasmania, 1995). Similarly, the proceedings of the AAMT conference *Springboards into numeracy* suggests that "being numerate empowers people and allows them to be critical users and developers of mathematics within specific contexts" (AAMT, 2002).

To refine our understanding of the first digits of numbers we find in everyday life we need to examine the numeracy of Benford's Law. I suggest here that it is not Benford's Law that is the issue; it is our understanding of the integers along with our ideas about randomness that causes the problem. We have adopted a mathematical picture of the integers and what they mean rather than one that conforms to the use of numbers in life.

Critical numeracy

If we define numeracy as the use of mathematics in real life situations, as before, then critical numeracy recognises that the use of mathematics in real life is socially constructed and reflects particular ways of thinking. Critical numeracy "is a focus on how numeracy in all its forms is involved in our relationships to each other and the world" (Stoessiger, 2002).

From a critical numeracy point of view it seems that we expect the first digits of numbers to be random in the way lottery numbers are random. We expect each first digit to occur $\frac{1}{9}$ of the time. Because we expect this we have not given much attention to the numbers we use and hence are surprised when a different distribution is found. Perhaps we have adopted a narrow mathematical view of numbers uncritically. We see no reason to believe that any individual integer is any more privileged than any other. We see no reason why 1 should be more used as a first digit than the others. Whereas our actual use of numbers in the world in many cases does not conform to this expectation. How is it that we have not noticed that nearly half the numbers (47.7%) we use start with a 1 or a 2?

I will discuss these issues more in the section *The critical numeracy of Benford's Law.*

Applications

The unexpected and counterintuitive nature of Benford's Law is exactly what has lead to its major application: the detection of fraudulent data. One of the reasons why we know that most people believe that first digits are randomly distributed is because when fraudsters invent data they ensure that their numbers start with a roughly equal number of the different digits. However if data does not approximate to Benford's Law it is likely to be fraudulent.

Mark Nigrini (1992, 1996, 1999) has done most to turn this into an efficient fraud detection procedure used by many corporations and tax agencies. It has been used to detect concocted tax returns, false accounting data, to scrutinise election results and to check for errors in data. The Law is used in computer

design to recognise that numbers most often start with 1 and hence it makes sense to optimise the processing of this digit. It may also be used to check the data produced by computer models to see if it is likely to conform to actual circumstances. It has also been used to check eBay auctions to see if they are rigged (Giles, 2007) and to check for price manipulation after the introduction of the euro as a European currency (el Sehity, Hoelzl & Kirchler, 2005).

It seems that Benford himself was not scrupulous in his tabulation of data but rounded the proportions so that the totals for each of his data sets summed to exactly 100%. Diaconis and Freedman (1979) showed that the probability of this happening by chance is astronomically small (p. 363). It is perhaps ironic that a forensic test should catch Benford in its net.

With the use of Benford's Law for fraud detection the wheel has turned a full circle; numerical data satisfying Benford's Law is regarded as proper while data not fitting the Law is seen as anomalous! But this very usage depends on most people not understanding that digits are distributed by Benford's Law. This nicely illustrates the critical numeracy contention that the use of mathematics is socially constructed.

Scale invariance

A major step forward in the understanding of Benford's Law came when Roger Pinkham (1961) showed that Benford's Law is scale invariant. In other words, multiplying the numbers by a scale factor or converting from pounds to kilograms or from one currency to another does not change the distribution of digits given by Benford's Law. He also showed that Benford's Law is the only distribution of digits which is scale independent so it is tempting to conclude that, in real world situations where the choice of units is arbitrary, it is the only candidate for explaining the observed distribution of digits. For example a physical phenomenon such as the size of rivers should not depend on the units used so perhaps the first digits should also be independent of those units. However much data does not obey Benford's Law and hence is not scale invariant. For example, a great deal of physical and man-made phenomena are random and random data is not scale independent under Benford's Law, nor is normal data such as people's heights or IQ, nor is linear growth such as simple interest. Not all unit changes are invariant under Benford's Law. Changing temperatures scales is not invariant for the initial digits. For example, changing any Celsius temperature to Absolute involves adding 273 so a collection of temperature between 10C and 1000C which follow Benford's Law will have 2 and 3 as first digit after conversion to Absolute temperatures and will not fit the Law.

In the real world, a scientific process rather than a mathematical one is required to decide what is actually happening. The mathematically most elegant formulations (e.g., Euclid's theorems) may be only approximations to how our world works. We need to investigate both the numeracy and mathematics of Benford's Law.

Random samples of random distributions

Our understanding of Benford's Law increased markedly when Hill (1988) showed that if distributions of data are selected at random (the distributions could be anything such as molecular weight tables, a country's tax data or sport statistics) and then random samples are taken from these distributions then the results follow Benford's Law. In other words, random samples of a random variety of distributions (be they lottery results, share prices, telephone numbers, company financial data, etc.) produce the distribution described by Benford's Law. A double random process is required. The initial distributions can be quite un-Benford like in their behaviour but if samples are taken from them in an unbiased manner and a variety of such distributions are sampled the resulting data obey Benford's Law. It should be noted that Hill requires the sampling to be done in a scale unbiased way. This result can explain why many collections of data such as in the pages of newspapers fit Benford's Law.

Hill (1993) suggests that while a single distribution selected at random may well be scale dependent by selecting from lots of different distributions (making sure that the selections are unbiased as to scale dependency) the results will cancel out any scale dependency in a single selection resulting in a scale-independent result. As Benford's Law is the only scale independent distribution of the leading digits the random samples from random distributions will converge towards Benford's Law. However there must be more to Benford's Law than this because there is much real world data which cannot be described as samples of distributions (for example single distributions) that fit the Law. In addition how can we be sure that the samples from real world data are unbiased with respect to scale dependency? Why would we expect newspapers which report almost exclusively in one currency and a single unit system, for example, to sample in such an unbiased way? There is a need to understand a lot more about how numbers and their leading digits come to us in our everyday and working lives.

Using entropy principles to derive Benford's Law

In an intriguing paper Kafri (2009) explicitly derives Benford's Law using a balls and boxes model. He related his results to Benford's Law by assuming that the single unit, that is, a 1, corresponds to a single ball and a digit is a box containing a number of balls. A sequence of such boxes makes a number. Kafri says, "It is assumed that, counter to common intuition (that the digits are the logical units that comprise numbers) that the logical units are the 1's. For example the digit 8 comprises of 8 lots of unit 1 etc." So instead of assuming the 9 digits are equally distributed among the possible nine first digit boxes (as in a random distribution) he varies both the units (i.e., the balls) and boxes randomly. The unexpected result is that the balls and hence the digits are not evenly distributed between the boxes. When the results are normalised to give a probability distribution, the surprising result is that the distribution is independent of the number of boxes: it depends only on the

number of balls. Applying his results to the digits in a given base (b) 1, 2, 3, …, b Kafri derived the expression for the probability p of integer n as

$$p(n) = \log_b \left(1 + \frac{1}{n}\right)$$

This, of course, is Benford's Law where n is a digit from 1 to $b - 1$ in the base b.

For this result to apply the first digit 1 is equivalent to one ball in a box, digit 2 to two balls and digit 9 is equivalent to nine balls. In other words, for Benford's Law to hold the first digits are not just integers but are quantities. Hence it is easier to get a 1 in a box through random events as a single ball is required but much harder to get a digit 8 because eight balls are needed. If the first digit is not just a random digit but is actually a quantity then their distribution will follow Benford's Law.

Other situations

There are some simple requirements for data to obey Benford's Law. Firstly the data must span at least one order of magnitude. Otherwise not all leading digits will be represented and hence the data cannot fit Benford's Law. Fewster (2009) has shown that the more orders of magnitude the better as any irregularities in the distribution can be averaged out over the different orders of magnitude. Data produced by chance processes on the integers such as lotteries will not follow Benford's Law because each of the nine digits will be equally represented—but lottery jackpot prizes do obey the Law (Fewster, 2009). Benford's Law does seem to be closely related to random processes but not the simple randomness of the lottery.

Nominal numbers, where numbers are used instead of labels or simply as place holders and which do not count, order or measure anything, will not obey Benford's Law. Telephone numbers clearly do not follow the Law. Neither will postal codes, social security numbers or license numbers. Data which is generated by the addition of a constant amount will not obey Benford's Law. It is easy to demonstrate that compound interest follows Benford's Law by setting up a table showing how an initial amount grows under compound interest. Collecting first digits demonstrates Benford's Law. However simple interest in which an initial sum grows by a constant amount does not fit the Law. Similarly, we would not expect temperature data which is changed between Fahrenheit, Celsius and Kelvin temperatures, all conversions which require adding a constant, to follow Benford's Law when expressed in the different units. Distributions of numbers such as normal distributions which cluster about a central number do not follow Benford's Law because the central numbers dominate. Hence shoe sizes and people's heights do not follow Benford's Law. As mentioned Benford's Law is the only distribution which is scale invariant and base invariant. Conversely data which is not scale or base invariant will not obey Benford's Law.

Google

Today Google with its ability to be used to search the numbers on the World Wide Web (WWW) is the repository of virtually all public numbers. Certainly there are private data sets which are not available to Google but virtually all numbers which have been published in any way will appear on Google. It is the super distribution of numbers used publicly. So Google provides an interesting test for Benford's Law. By collecting all distributions it could be expected to be free of scale bias and hence the numbers should fit Benford's Law.

There have been several studies of numbers indexed by Google. Solomon (2006) put a thousand random six-digit numbers into Google and obtained a Benford like distribution. Brian Silverman (2003) sent consecutive integers to Google in August 2003 and recorded the number of hits. From his graph of the first 10 000 results he noticed a nested structure with major peaks (many more hits) on the powers of ten. Numbers starting with the digit 5 recorded many more than surrounding numbers. Not surprisingly there was one huge peak at 2003. He did not analyse his results for Benford's Law like behaviour but it is clear that, despite the logarithmic nature of his results, the major spikes in the data make it unlikely the Googled numbers would fit Benford's Law closely.

Greg Leibon (2008) developed a sample project for his students on Google numbers. He wished to investigate numbers on the WWW "in which real live people were actually interested", as distinct from collections intended only for data mining. To accomplish this, he studied six-digit numbers and included a piece of text with his search so that the search only included numbers which were accompanied by text. For text he used the word "nature". He constructed his sample of six-digit numbers by randomly generating nine five-digit numbers and then put the digits 1 to 9 in front of each.

The results are shown in Figure 1, revealing an excellent fit to Benford's Law. Leibon (2008) confirmed this with a chi squared test. Including some text with the data, eliminating large data mining collections, seems to bring the WWW numbers searched by Google into much better agreement with Benford's Law. Leibon examined the Google results of some of the numbers he had used and noticed that many of them were related to growth. He concluded that WWW numbers (with text) are largely following Benford's Law because they are from growth data.

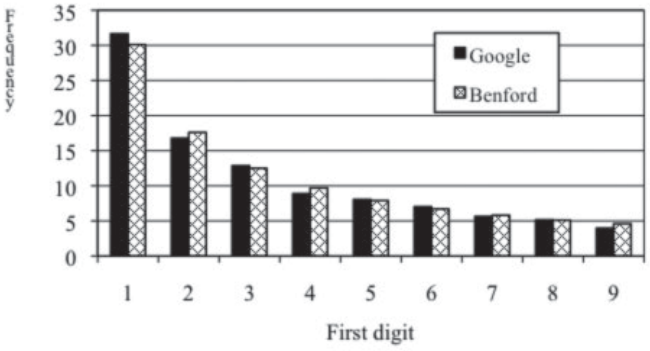

Figure 1

Dorovgetski, Mendes, and Oliveira (2005) have studied numbers on the WWW. They too noted the very high frequency of occurrence of the current date. They found that numbers on the WWW fell into a number of different classes. For example, powers of 10 occur with "strikingly high frequency" (p. 552) and formed a class of their own with a different distribution to non-round numbers. The very high frequency of occurrence of the powers of 10 seems to be due to our base 10 system and hence is obviously not base independent. Certainly Benford's Law does not apply to the WWW numbers as a whole.

In summary, WWW numbers accessed through Google as a totality do not seem to be a good fit to Benford's Law. The very high incidence of powers of ten is one factor which biases the numbers. Ironically this can lead to an over representation of the digit 1 and hence Benford like distributions. The prevalence of powers of 10 suggests that the base of our number system is important—this is hardly surprising. This implies that we cannot assume that real world data collections are base invariant and thus conform to Benford's Law.

Growth data

We live in an expanding commercial and financial world. Companies grow, economies expand, currencies inflate and incomes (generally) rise. It is possible that much of the data reported about the world is growth data and that growth data may be linked to Benford's Law.

Linear growth

Linear growth, or growth by a constant amount such as simple interest, does not lead to a Benford's Law type distribution. Because the numbers are increasing in constant increments it takes just as many steps to get from 1 to 2 (100 to 200, etc.) as it does to get from 2 to 3 (200 to 300), and so on. In simple interest terms, if the interest rate is 5% then it takes 20 time units to get from \$100 to \$200 and 20 time units to get from \$800 to \$900 so each starting digit

is equally likely. Constant growth cannot explain Benford's Law. This is not the case for exponential growth, however.

Exponential growth

Much growth is exponential. As a company or economy grows, it mobilises new resources which allow it to grow even faster. Compound interest is an obvious example of exponential growth.

We can show the relationship between exponential growth and Benford's Law more formally. The compound interest formula is $a = p(1 + i)^n$ where a is the accumulated amount, p is the principal, i is the interest rate and n the number of time periods. This gives the number of time periods for growth from p to a as

$$n = \frac{\log\left(\frac{a}{p}\right)}{\log(1+i)}$$

So for a given interest rate the number of time periods to change from p to a is proportional to $\log\left(\frac{a}{p}\right)$.

The growth has to be at least an order of magnitude if it is to cover all nine leading digits (and hence for Benford's Law to apply) so we can take an arbitrary starting point at a power of ten and finish on the next power of ten. So the principal p can be represented in terms of the first digit D as $p = D(10^z)$. When growth results in the accumulated amount reaching the next digit, $D + 1$, then $a = (D + 1)(10^z)$. The number of time periods to go from D to $D + 1$ will be proportional to

$$\log\left(\frac{D}{D+1}\right)$$

Hence the periods of time for which an account (company) et cetera will report digit D relative to digit $D + 1$ will be

$$n = \log\left(\frac{D+1}{D}\right) \text{ or } n = \log\left(1+\frac{1}{D}\right)$$

This is Benford's Law stated in terms of the number of time periods it takes to change from first digit D to $D + 1$ rather than the probability that the first digit is D.

In this perspective Benford's Law is simply about the time periods taken for numbers to grow from one leading digit to the next while compounding at a constant growth rate. Under compound growth rates it simply takes longer to get from 1 to 2 than from 2 to 3, etc., as specified by Benford's Law.

In exponential growth the initial digit is important. Numbers in the 900s (9000s etc.) grow much more quickly than those in the 800s. Numbers in the 200s grow nearly twice as fast as those in the 100s. Hence any data reporting growth has many more of the slower growing digits, that is, the ones and twos as the leading digit than the faster growing eights and nines. This is true for any similar exponential function.

Is the occurrence of exponential growth enough to explain the observations of Benford like distributions in situations such as company accounts, stock exchange prices or auctions where the explanation in terms of random samples from random distributions does not apply? To some extent it is. Certainly where populations or financial resources are compounding steadily Benford's Law will apply. However while Benford's Law may apply very well to steadily growing populations there is reason to believe it might not be so reliable for financial data reported by individuals, companies or countries. For example while companies that survive usually grow over time they may grow for a while then spend some of their profits on expansion or return dividends to share holders. They may stagnate for some time or make a loss now and then. Steady compounding growth seems too strong a model to account for the numerical data produced by such concerns. Are there other ways of explaining such situations?

Growth data: Another explanation

Mark Nigrini who pioneered the use of Benford's Law as a forensic tool for company accounts and taxation data suggested the following explanation of how the Law works.

He was quoted in the *New York Times* (Browne, 1998) as explaining that as a stock exchange average, such as the Dow Jones, grows from 1000 to 2000 the average increase is 100%. If it increases by a rate of 20% it will take five years to get from first digit 1 to first digit 2. During that time the first digit of the average will always begin with 1. But if the average starts at 5000 it will only be a 20% increase required to get to 6000 and that increase will occur in only one year. Similarly when the average reaches 9000 it will only take an 11% increase which will happen in seven months. Nigrini explains: "As you can see, the number 1 predominates at every step of the progression, as it does in logarithmic sequences" (p. 5).

Following Nigrini's explanation it is interesting to speculate that the leading integer data might be explained by the simple idea that it requires 100% growth to go from 1 to 2 (or 1000 to 2000, etc.) but only 50% growth to go from 2 to 3 (or 2000 to 3000, etc.), and so on. This leads to the results in Table 2.

If these fractions represent the ratios of the number of time steps required to achieve the next initial digit, and hence the number of times a particular digit will be reported, they should be related to the probabilities

Table 2. Fractional growths from one digit to the next.

Leading digit	Fractional growth
1	$\frac{1}{1}$
2	$\frac{1}{2}$
3	$\frac{1}{3}$
4	$\frac{1}{4}$
5	$\frac{1}{5}$
6	$\frac{1}{6}$
7	$\frac{1}{7}$
8	$\frac{1}{8}$
9	$\frac{1}{9}$
Total	2.83

that the digits will occur. This can be most simply done by normalising the probabilities by dividing by the sum of the fractional growths, 2.83. The results are shown in Table 3 alongside the probabilities given by Benford's Law for comparison. The distributions are similar but by no means the same.

Table 3. Fractional growth probabilities compared to Benford's Law.

Leading digit	Fractional growth	p(fractional growth)	p(Benford's Law)
1	$\frac{1}{1}$	0.353	0.301
2	$\frac{1}{2}$	0.177	0.176
3	$\frac{1}{3}$	0.118	0.125
4	$\frac{1}{4}$	0.088	0.097
5	$\frac{1}{5}$	0.071	0.079
6	$\frac{1}{6}$	0.059	0.067
7	$\frac{1}{7}$	0.050	0.058
8	$\frac{1}{8}$	0.044	0.051
9	$\frac{1}{9}$	0.039	0.046
Total	2.83	1.00	1.00

Zipf's Law

The distribution in column 3 of Table 3 is actually well known. It is a version of another important growth law, Zipf's Law (1949), originally formulated to account for the frequency distribution of words in the English language. Zipf's Law was formulated by the American linguist George Zipf from his studies of the occurrence of words in text. It states that if the frequency of words are counted in a body of language then a word that is ranked nth in frequency has its probability of occurrence proportional to $\frac{1}{n}$. Hence if 'the' is the most commonly occurring word and 'of' is the next most common word then 'of' will occur $\frac{1}{2}$ of the number of times that 'the' occurs. Zipf's Law has now been widely applied to a variety of other statistical data such as population, company size and geographical data. It can be extended by including additional variables to give a better fit to the actual data. However it is not often used as in Table 3 where the distribution is normalised across a limited number of possible outcomes to give the probabilities of those outcomes. Hence Table 3 is the application of Zipf's Law to the leading digits 1 to 9 given their rank order is also 1 to 9 and no other digits are possible.

Of course Zipf's Law would usually extend down the integers, 11, 12 and so on. However there are only nine leading digits so in this case it is necessary to consider these nine on their own. The Law can be extended to two digit numbers as a set of their own, similarly for three digit numbers et cetera. In these cases there is very little difference between Zipf's and Benford's Laws (Bogomolny, n.d.). The major differences between the two formulations are that Zipf's Law predicts a frequency for digit 1 which is 17% greater than does

Benford's Law with correspondingly lower frequencies of digits 7, 8 and 9. A graphical comparison between the two Laws is shown in Figure 2.

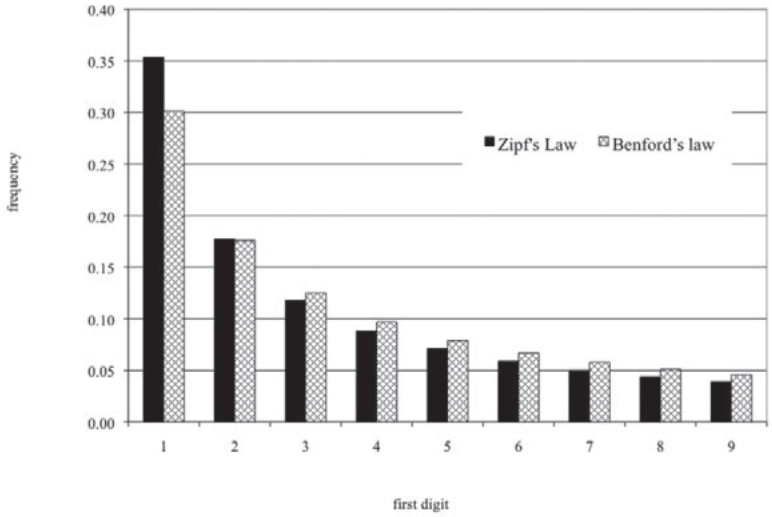

Figure 2. Frequencies of first digits given by Zipf's Law and Benford's Law.

As mentioned above, Benford's Law is the only distribution that is scale-independent so if this is required of a data set then it is the only candidate. But much real world data depends on the units it is expressed in. So if scale independence is not required of a data set then we have two possible explanations: Benford's and Zipf's. They are very similar the most obvious difference being in the frequency of the first digit 1 with Zipf's Law predicting it occurring at a higher frequency than Benford's Law.

Now much real world data is about growth. Companies grow (if they decline for too long they vanish), economies grow (with occasional decline during recessions), cities grow, populations grow, auction bids grow and stock prices grow. Which of Benford or Zipf's Laws best model such data? As we have said before, Benford's Law exactly fits compound interest while simple interest fits neither.

What does Zipf's Law represent?

Zipf's Law for the nine initial digits models a different form of growth from constant exponential growth. It is intermediate between simple and compound interest. In practical growth terms it is equivalent to a company growing in a constant way (simple interest) from one leading digit to the next (e.g., from 100 to 200), then compounding the growth at this digit level and then growing with constant growth again until the next digit is reached (from 200 to 300). The simple interest formula is $A = P(1 + in)$ where A is the accumulated amount , P the principal, i is the interest rate (as a fraction) and n the number of time periods. This gives the number of time periods for growth from P to A as

$$n = \frac{A - P}{iP}$$

Expressing P as $P = D(10^z)$ as before, the growth from $P = D(10^z)$ to the next digit, $D + 1$, gives $A = (D + 1)(10^z)$. So

$$n = \frac{10^z}{i \times D \times z} \text{ or } n = \frac{1}{iD}$$

So the number of time periods to grow by simple interest from initial digit D to the next digit $D + 1$ for a given interest rate is proportional to $\frac{1}{D}$. Instead of continuing on with interest calculated on the initial principal we now calculate the simple interest starting at the next digit amount. This can be described as compounding the principal at this new digit level and then simple interest is applied again to the new principal The number of time periods will again be proportional to the inverse of the new digit, $\frac{1}{D+1}$. This is Zipf's Law for the first digits. Hence Zipf's Law is equivalent to a process in which simple interest is applied to grow from an initial first digit until the next one, the amount is compounded at this digit level and simple interest is applied again and so on.

This may be a good model for growth in certain situations. For example a company may grow by a constant amount for a while then may consolidate by the purchase of new plant or equipment, buying a rival or expanding in a new market. After the consolidation it may grow steadily again but on the expanded base of the new plant or equipment until it builds up the resources for further consolidation. In such a situation the company's results will be represented better by Zipf's Law than by Benford's Law. A detailed study of actual data would be required to distinguish between the various possibilities. Given the differences between the predicted frequencies of the digits 1,7, 8 and 9 in the two theories it may be possible to reach conclusions about particular data sets.

The critical numeracy of Benford's Law

We have seen that Benford's Law has been described as both mysterious and ubiquitous. There must be something wrong. It is probably our understanding of the integers which is faulty.

The way we think about the nine integers which we use as first digits and our social constructions about them are largely unexamined. We almost take them as god given. In fact the 19th century mathematician Leopold Kronecker (Bell, 1986) made the observation: "God made the integers, all else is the work of man". The integers, he believed, were handed down to us and then used to develop all the other numbers and the rest of mathematics. But our misconceptions about first digits suggest we need to look at them more closely.

WWW data as searched by Google should contain virtually all public data but these show many deviations from Benford's Law. It is not base independent but shows a predominance of numbers that are powers of ten reflecting our base 10 system. However, if the numbers accessed by Google are sampled by including some text alongside them then Benford Law like behaviour results.

Benford's Law can be explained as numbers resulting from growth processes. Benford's Law corresponds to exponential growth such as compound interest. Simple interest does not fit Benford's Law. A Zipf's Law model for the first digits corresponds to simple interest growth from one leading digit to the next which is compounded at this digit then grows simply again. This may be a better representation of reality for many businesses.

Benford's Law can be explained in terms of the random distribution of balls in boxes. In this model the integer 1 corresponds to one ball, 2 to two balls, etc. It is a lot harder to randomly get nine balls in a box than one and this gives rise to the Benford distribution. Hence although we think of the nine digits (plus zero in later places) as the units which make up our number system this may be a misunderstanding. If the only unit is 1 and the other digits are composite then Benford's Law can arise from random distributions of the unit.

Perhaps Kronecker's claim is too strong! Maybe we have been given only the digit 1. From this digit the other digits are constructed and then are used along with the place holder zero to represent the integers in a given base system. This then gives the most straight forward understanding of Benford's Law and how numbers are represented in the world. First digits are not equally likely because the digits themselves are not equal. The higher digits are collections of the unit digit and hence it is less likely that several unit digits are collected to form a larger digit.

In addition our ideas about randomness may be inadequate. In statistics we usually equate random with 'equally likely'. This is the randomness of the lottery where each number is equally likely to be drawn. This seems to be how we think of randomness in general but Benford's Law tells us that these ideas do not conform to reality in general. However thermodynamic randomness has a different quality—all the different states of, say, balls in boxes are random. The numbers of balls are not equally likely just the different micro-states. In this situation, Benford's Law applies and this seems to be how leading digits are distributed in many situations in our world.

The critical numeracy perspective is that we have a faulty idea of the integers. Yes, they can be random but in practical situations they are often not. If they are used to measure something, such as money, then first digit nine is very different from first digit one because nine requires the accumulation of nine ones and will be much harder to achieve.

Conclusions

A critical numeracy examination of Benford's Law suggests that our understanding of the integers is faulty. We think of them as equally likely to turn up as the first digit of a random real world number. For many real world data sets this is not true. In many cases ranging from eBay auction prices to

six digit numbers in Google to the distribution of numbers in newspapers the smaller digits are much more likely than the larger ones. Yet most of us are surprised when we first encounter this result. Benford himself described the real world numbers which fit his law as anomalous. Many others have echoed his surprise. How can our understanding of numbers be such that the way we actually use numbers in our world, that is, the authentic use of numbers, is regarded as strange?

The distribution of numbers in this way has been explained in the past by Benford's Law. However it seems that Zipf's Law may be just as useful as an explanation of some of the observed distributions. Both laws are likely to apply when numbers describe growth situations with Benford's Law describing compound interest type growth while Zipf's Law represents a slower growth with constant growth (simple interest) compounded at repeated stages.

From a critical numeracy perspective we need to understand how the first digit distribution of real world numbers is both ubiquitous but seen as anomalous.

Perhaps this is best explained using the work of Kafri (2009) showing that the random distribution of balls and boxes results in a Benford's Law distribution. In this model the digit 1 is represented by a single ball, digit 2 by two balls and so on. In this model the first digits are actually quantities of the single unit digit. Kafri used a thermodynamic randomness to distribute the balls (rather than a lottery style) and Benford's Law is the predicted distribution. Hence our understanding of random as meaning equally likely is too simplistic for real world numbers. The digits in numbers are not distributed as if by lottery. We need to move to a thermodynamic understanding of randomness in which we recognise that digits are a quantity of the unit digit and are distributed amongst all the different positions which make up a number. It is all the different micro-states of digit as quantity in all the various positions which are 'equally likely'.

A better understanding of the use of numbers in our world seems to be that:

1. Some numbers such as lottery results are random and their first digits are uniformly distributed.
2. Some numbers represent quantities (such as amounts of money) and it is harder to accumulate say, $700 than $100. In these situations the digit 1 is the basic unit and the other digits are collections of a number of ones and are less likely (in a thermodynamic sense). The distributions of the first digits of these numbers are well described by Benford's Law.
3. The initial digit of numbers from random samples taken from a random variety of distributions will fit Benford's Law.
4. The super sample of numbers held in Google is not particularly Benford like. The dominance of the base ten number system results in large spikes on the powers of ten. Our use of numbers is not base independent—hardly surprising—but it needs to be recognised as a limitation to the use of Benford's Law.

5. Growth data resulting from regularly compounding growth follows Benford's Law. However this may be too strong for many real world situations where Zipf's Law may better represent less regular growth.

In addition, a critical numeracy examination suggests that our understanding of numbers is deficient in that on the one hand we know that the digit 9 is composed of nine ones but we do not expect numbers to behave that way when we use them in the world. Perhaps we have been conditioned by lottery results and other similarly random number data sets to expect an 8 or 9 to turn up as often as a 1. Real world numbers are not, in general, like that. You have to work a lot, lot harder to accumulate 8 million dollars than 1 million.

References

The Australian Assocition of Mathematics Teachers [AAMT]. (2002). *Springboards into numeracy*. Retrieved from http://www.aamt.edu.au/Activities-and-projects/Previous-projects/Focus-conferences/Springboards

Bell, E. T. (1986). *Men of mathematics*. New York: Simon & Schuster.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society, 78* (4), 551–572.

Bogomolny, A. (n.d.). *Benford's Law and Zipf's Law*. Retrieved from http://www.cut-the-knot.org/do_you_know/zipflaw.shtml.

Browne, M. W. (1998, August 4). Following Benford's Law, or looking out for No. 1. *The New York Times*, p. 5.

Diaconis, P., & Freedman D. (1979). On rounding percentages. *Journal of the American Statistical Association, 74*, 359–364.

Dorogovtsev, S., Mendes, J., & Oliveira, J. (2005). Frequency of occurrence of numbers in the World Wide Web. *Physica A, 360*, 548–556. http://arxiv.org/pdf/physics/0504185

Education Department of Tasmania. (1995). *Numerate students: Numerate adult*. Hobart: Author.

el Sehity, T., Hoelzl, E., & Kirchler, E. (2005). Price developments after a nominal shock: Benford's Law and psychological pricing after the euro introduction. *International Journal of Research in Marketing, 22*, 471–480.

Fewster, R. M. (2009). A simple explanation of Benford's Law. *The American Staistician, 63*, 26.

Giles, D. E. (2007). Benford's Law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters, 14* (3), 157–161. Retrieved from http://web.uvic.ca/econ/ewp0505.pdf.

Hill, T. P. (1988). The first digit phenomenon. *American Scientist, 86*, 358–363.

Hill, T. P. (1993). The difficulty of faking data. *Chance, 17*, 30.

Kafri, O. (2009). *Entropy principal in direct derivation of Benford's Law*. Retrieved from http://arxiv.org/ftp/arxiv/papers/0901/0903047.pdf, 2

Leibon, G. (2008). *Google numbers*. Retrieved from http://www.dartmouth.edu/~chance/chance_news/for_chance_news/ChanceNews13.03/GregProject.pdf

Ley, E. (1996). On the peculiar distribution of the US stock indices' digits. *American Statistician, 50*, 311–313.

Mathews, R. (1999). The power of one. *New Scientist* July 26. Retrieved from http://www.fortunecity.com/emachines/e11/86/one.html

Newcomb, S. (1881). Note on the frequency of the use of digits in natural numbers. *American Journal of Mathematics, 4*, 39–40.

Nigrini, M. J. (1992). *The detection of income tax evasion through an analysis of digital frequencies* (Unpublished PhD thesis). University of Cincinnati, Cincinnati, OH.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's Law. *Journal of the American Taxation Association, 18,* 72–79.

Nigrini, M. J. (1999). I've got your number. *Journal of Accountancy, 187*, 79–83 Retrieved from http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm

Pinkham, R. (1961). On the distribution of the first significant digits. *The Annals of Mathematical Statistics, 32*, 1223–1230.

Rami, R.A. (1976). The first digit phenomena. *American Mathematical Monthly, 83*, 52.

Silverman, B. (2003). *Googling for integers*. Retrieved from http://forum.wolframscience.com/showthread.php?threadid=116.

Solomon, M. (2006). *Demonstrating Benford's Law with Google*. Retrieved from http://www.thecleverest.com/benfords_law.html

Stoessiger, R. (2002). An introduction to critical numeracy. *The Australian Mathematics Teacher, 58* (4), 17–20.

Weisstein, E. (2010). *Benford's Law. From MathWorld—A Wolfram Web Resource*. Retrieved from http://mathworld.wolfram.com/Benfordslaw.html.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison Wesley.